Multinomial Processing Tree (MPT) Modeling, Part 3: Advanced Features of multiTree

Edgar Erdfelder and Daniel W. Heck

University of Mannheim & Phillips University Marburg

3) Advanced features of multiTree

- 3.1) Identifiability checks in multiTree
- 3.2) Statistical power analyses
- 3.3) Model selection
- 3.4) Summary: Model Evaluation

3.1) Identifiability checks in multiTree

- Get Jacobian
  - Evaluate at random parameter location
  - Use parameter values from parameter tab
- Repeated analysis
  - Check stability of parameters estimates for a specific vector of observed frequencies
- Simulated identifiability
  - Repeated data generation for different random locations in Ω, followed by estimation

3.2) Statistical power analyses

Traditional Power Analysis (Cohen, 1969, 1972, 1988)

- Under H_1, any Power-Divergence statistic (S_λ) follows a noncentral chi-square distribution with noncentrality parameter $\gamma_{(\lambda)} = N \cdot \mathbf{w}_{(\lambda)}^2$, where $\mathbf{w}_{(\lambda)}$ denotes the effect size $\mathbf{w}_{(\lambda)}^2 = S_\lambda(\Theta)/ N$
- Effect size conventions
  - w = .10 (“small effect“)
  - w = .30 (“medium effect“)
  - w = .50 (“large effect“)
- Types of power analysis
  - “A priori”: Compute N as a function of w, α, and 1-β
  - “Post hoc”: Compute 1-β as a function of w, α, and N

Problems of the traditional method

- Problem 1:
  - How does effect size translate into (differences between) parameter values?
- Problem 2:
  - Do effect size labels have the same meaning in different models?
- Problem 3:
  - The relative size of groups (conditions) is ignored.

Cohen (1988, p. 244)

On *w* effect sizes conventions:

"Their use requires particular caution, since, apart from their possible inaptness in a particular substantive context, what is subjectively the same degree of departure or degree of correlation (…) may yield varying *w*, and conversely. **The investigator is best advised to use the conventional definitions as a general frame of reference (…) and not to take them too literally**."

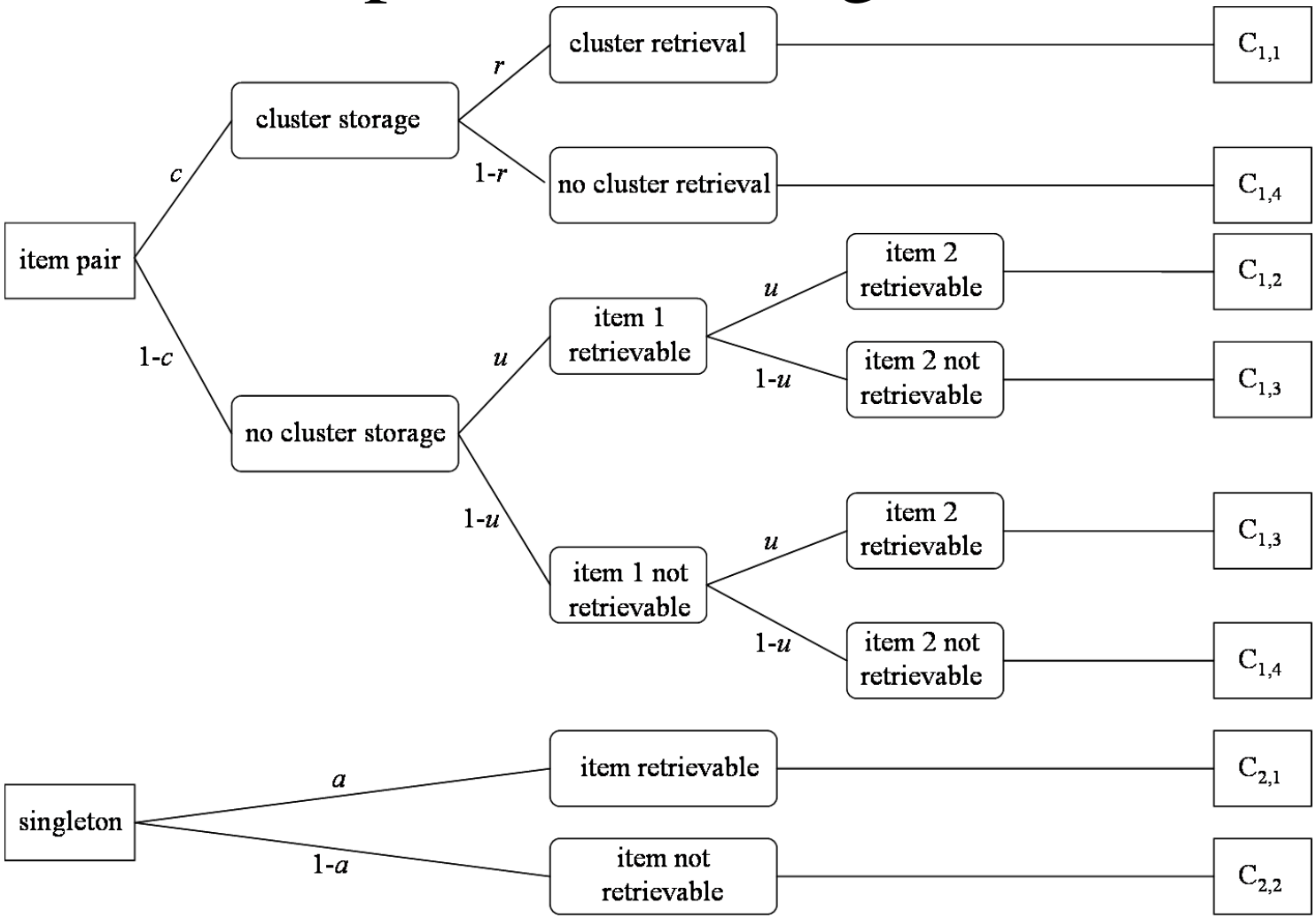
Approach 2: Power as a function of the model parameters under H_1

1) Specify your H_0 model
2) Specify your H_1 model with all parameter values fixed at „plausible values“
3) Choose N_k, $k = 1, ..., K$, and calculate expected frequencies under H_1
4) Fit the H_0 model to the H_1 expected frequencies by minimizing PD^{λ} for your choice of λ
5) Use the minimum PD^{λ} value as the noncentrality parameter $\gamma_{(\lambda)}$
6) Compute $1\text{-}\beta(\pi) = Pr(\chi^2(\gamma_{(\lambda)}, df) \geq c_{(df, \alpha)})$

Open questions

- How well does Approach 2 work for joint MPT models?
- How does approximation quality depend on the sample sizes?
- Does the approximation quality depend on the PD^{λ} test statistic used?

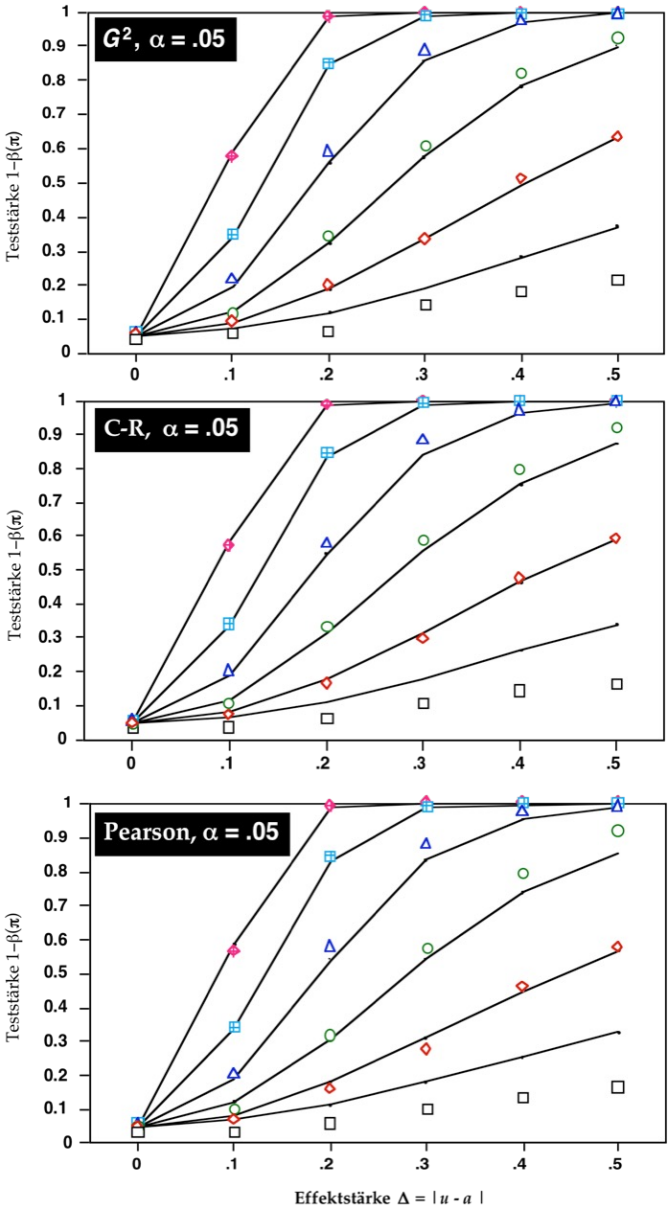
Let‘s return to our running example: The pair clustering model



A Monte-Carlo study

- Storage-retrieval model ($\tau_1 = 2/3$, $\tau_2 = 1/3$)
- H_0: $u = a$, $df = 4\text{-}3 = 1$
- H_1: $c = r = .5$, $u = .5 - \Delta/2$, $a = .5 + \Delta/2$, for
  $\Delta = .00 / .10 / .20 / .30 / .40 / .50$
  ($w = .00 / .07 / .13 / .19 / .24 / .28$)
- $N = 30 / 60 / 120 / 240 / 480 / 960$
- Type-1 errors $\alpha = .01 / .05$
- Statistics: G^2, X^2, Cressie-Read
- 1000 Monte Carlo samples per run

Results for $\alpha = .05$



- *Symbols*: Monte-Carlo estimated power for G^2 (top), Cressie-Read (middle), and X^2 (bottom).
- *Black lines*: Approximate power using same λ in ncp $\gamma_{(\lambda)}$ as in PD^{λ} test statistic
- Sample sizes from top to bottom:
  - $N = 960$
  - $N = 480$
  - $N = 240$
  - $N = 120$
  - $N = 60$
  - $N = 30$

Absolute differences between Monte-Carlo power and approximate power formula ($\alpha = .05$)

	Test statistic used					
	G^2		*C-R*		X^2	
N	mean	max.	mean	max.	mean	max.
30	.065	.157	.075	.173	.078	.162
60	.007	.019	.010	.020	.015	.032
120	.021	.040	.027	.049	.031	.066
240	.017	.031	.020	.041	.021	.048
480	.004	.011	.005	.010	.006	.014
960	.003	.014	.003	.014	.004	.016
Mean	**.020**	**.045**	**.023**	**.051**	**.026**	**.056**

Conclusions from the Monte-Carlo study

- In our example, the power approximation is acceptable for $N > 50$ and good for $N > 200$
- Approximation accuracy appears to be worse for $\alpha = .01$ compared to $\alpha = .05$
- Approximation accuracy is slightly larger for G^2 compared to both the Cressie-Read statistic and Pearson´s X^2.
- Hence, G^2 can safely be used as a goodness-of-fit statistics combined with $\alpha = .05$ and power analyses based on the parameter values under H_1 (as sketched above).
- Prior to each study, the N per tree of the MPT model should be determined that yields sufficient power for key χ^2 tests applied to model parameters.

Power optimization

Problem:

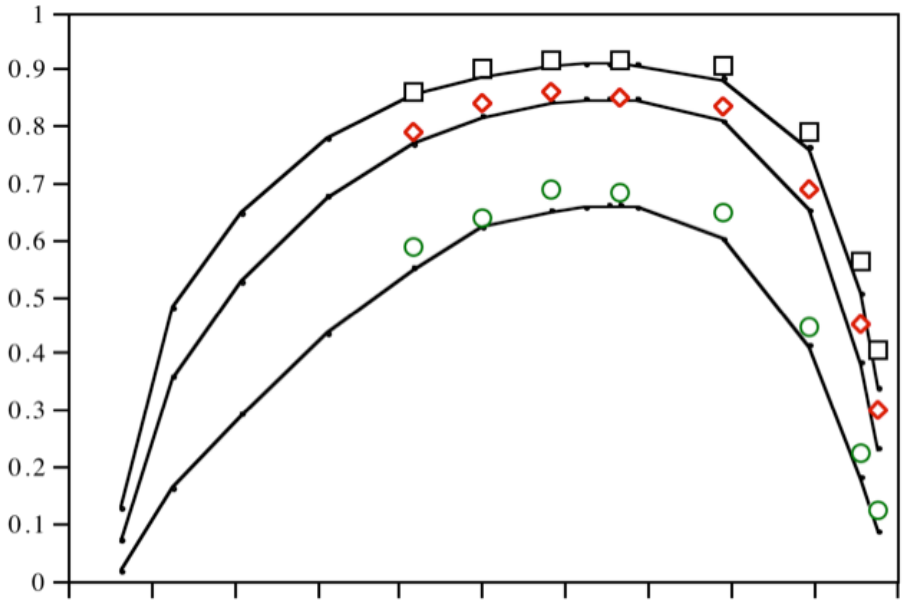
Given a fixed total sample size and a fixed α, is there any way to maximize the power?

Answer:

Yes, there is!

1) τ_k Optimization

- Power of G^2 as a function of τ_1 for the storage retrieval model, given α = .10 (top), .05 (middle) and .01 (bottom).
- $c = r = .50$, $u = .40$, $a = .60$
- $N = 480$
- Conclusions:
  - Strong effect of τ_1 !!
  - Max. power for $\tau_1 = .652$
  - Thus, $\tau_1 = ... = \tau_K$ may be a bad default option!



2) θ Optimization

- Model parameters can be divided in H_0-relevant and H_0-irrelevant parameters. For the pair-clustering model test:
  - *u* and *a* are H_0 relevant
  - *c* and *r* are H_0 irrelevant
- Problem:

  How to choose the values of the H_0-irrelevant parameters so as to maximize the power of the model test?

Approximate power of the G^2 test for the storage-retrieval model ($\alpha = .05, N_1 = 320, N_2 = 160$)

Parameter values under H_1				ncp	approx
c	r	u	a	$\gamma_{(0)}$	$1-\beta(\pi)$
.10	.80	.40	.60	12.49	.94
.10	.20	.40	.60	12.49	.94
.50	.80	.40	.60	8.92	.85
.50	.20	.40	.60	8.92	.85
.90	.80	.40	.60	2.53	.36
.90	.20	.40	.60	2.53	.36

3) Power of conditional versus unconditional tests

- Consider two nested models:
  - M_0 with parameter space Ω_0
  - M_1 with parameter space Ω_1
  - $\Omega_1 \subset \Omega_0$
- Problem:

  M_1 can be tested by an unconditional or a conditional G^2 test provided that M_0 holds. Which test is more powerful?

Running example: pair-clustering model

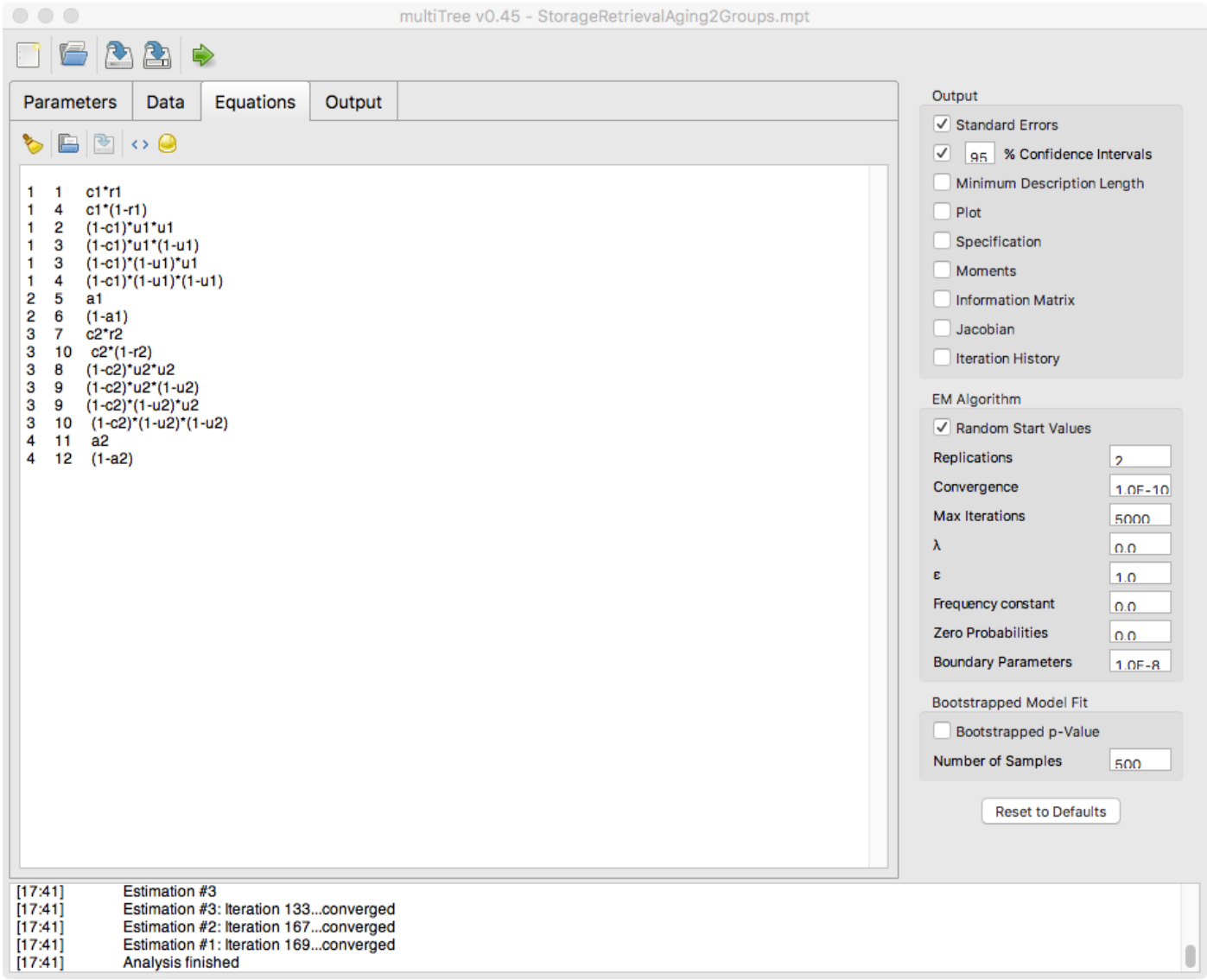
- $N_1 = 160, N_2 = 80$
- M_0: $u = a$
- M_1: $u = a$ and $c = .30$
- Under H_1: $c = r = u = a = .50$ we obtain ($\alpha = .05$):
  - $G^2(M_1)$: df = 4-2 = 2, $\gamma_{(0)} = 8.62$, 1-$\beta(\pi) = .75$
  - $G^2(M_1)$-$G^2(M_0)$: df = 2-1 = 1, $\gamma_{(0)} = 8.62$, 1-$\beta(\pi) = .85$
- Therefore, use conditional G^2 difference tests whenever possible as they are more powerful.

Summary and conclusions

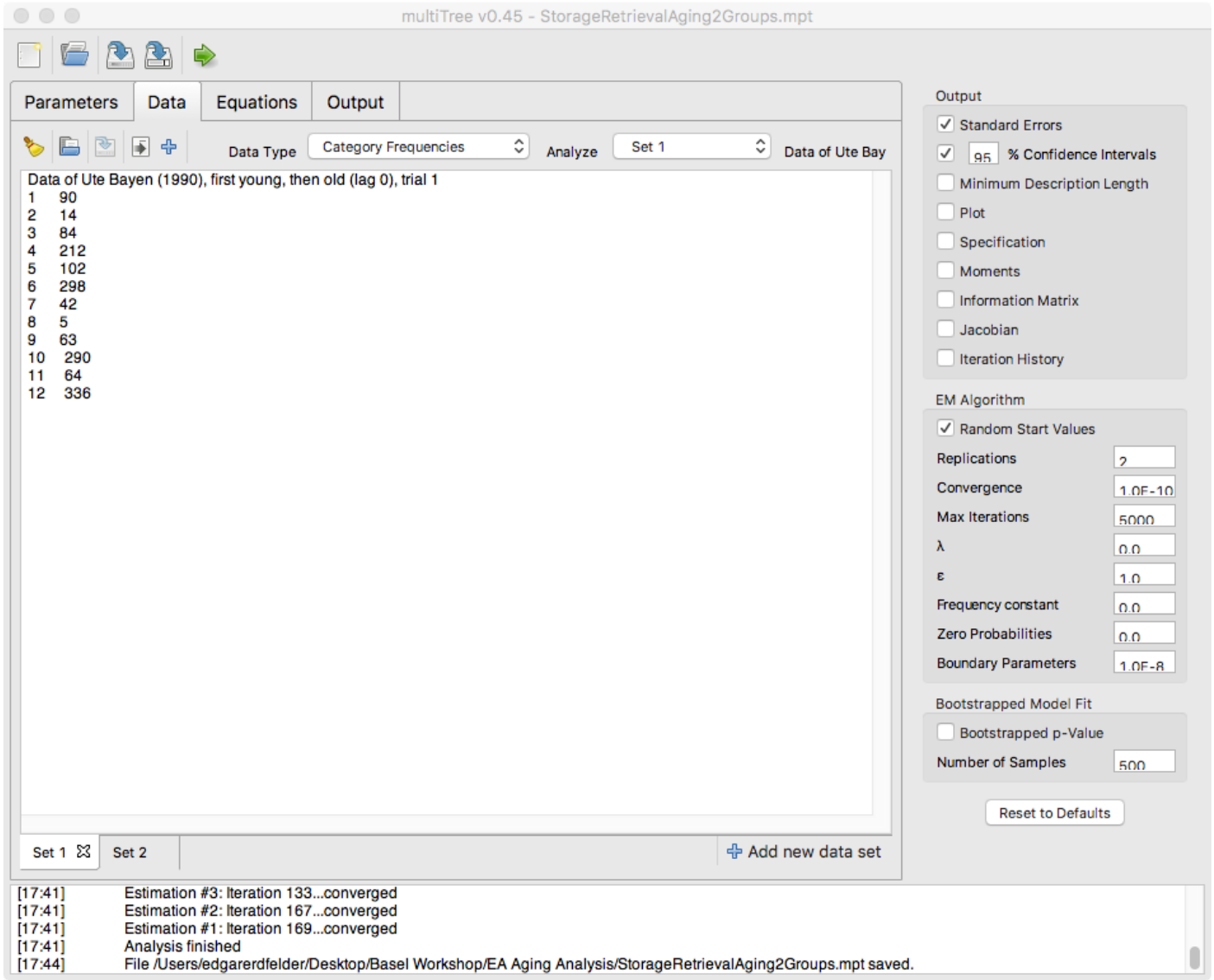
- Do not ignore the power of model tests!
- The proposed approximation method works very well for joint MPT models with typical sample sizes
- G^2 is a good default option for several reasons:
  - Approximation accuracy is optimal
  - Maximum power for “diffuse” noncentrality structures
  - Option of conditional tests
- Do not forget to optimize context conditions:
  - τ_k optimization
  - θ optimization
  - conditional tests whenever possible.

A priori and post-hoc power analyses using multiTree

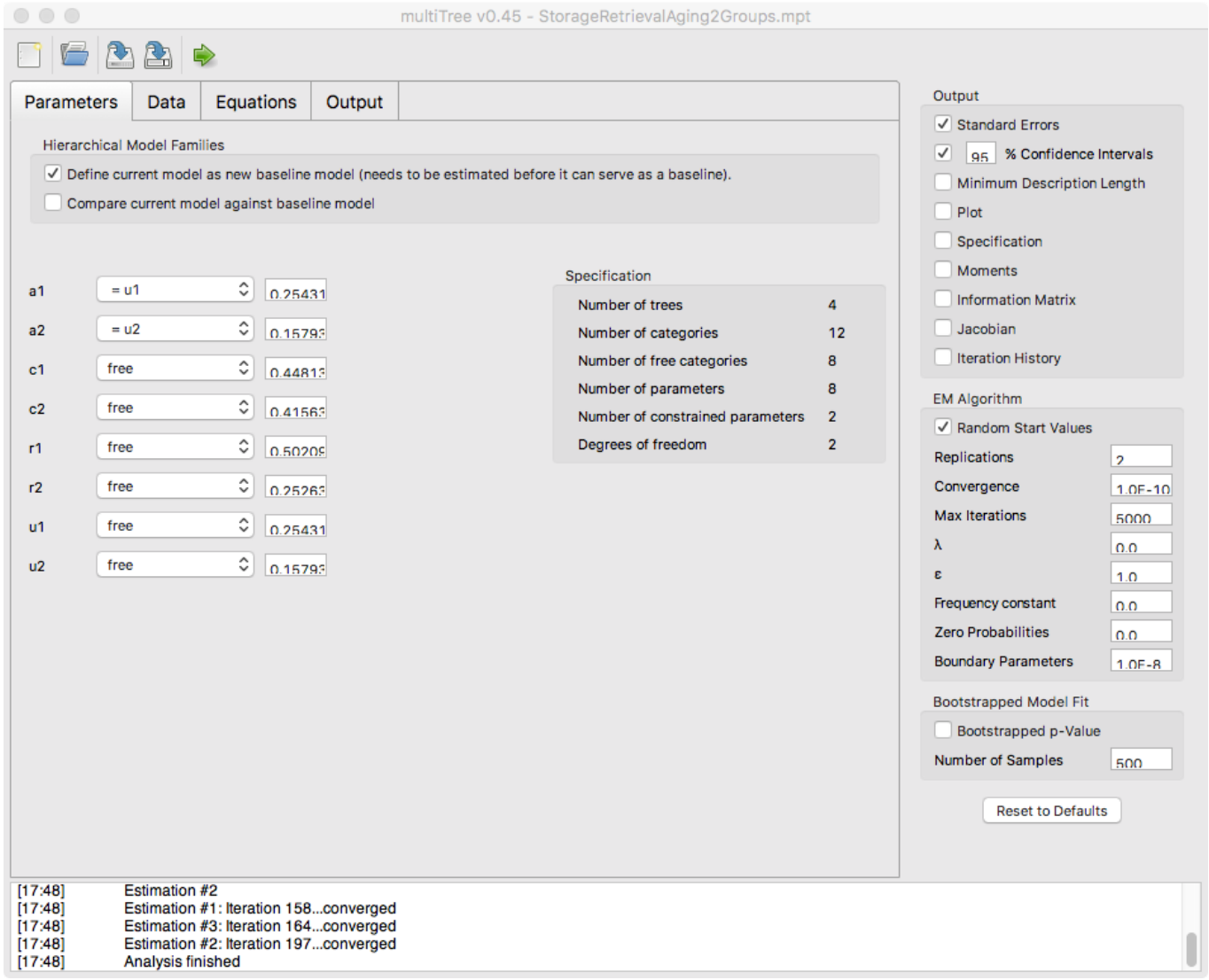
Example: Age group comparison



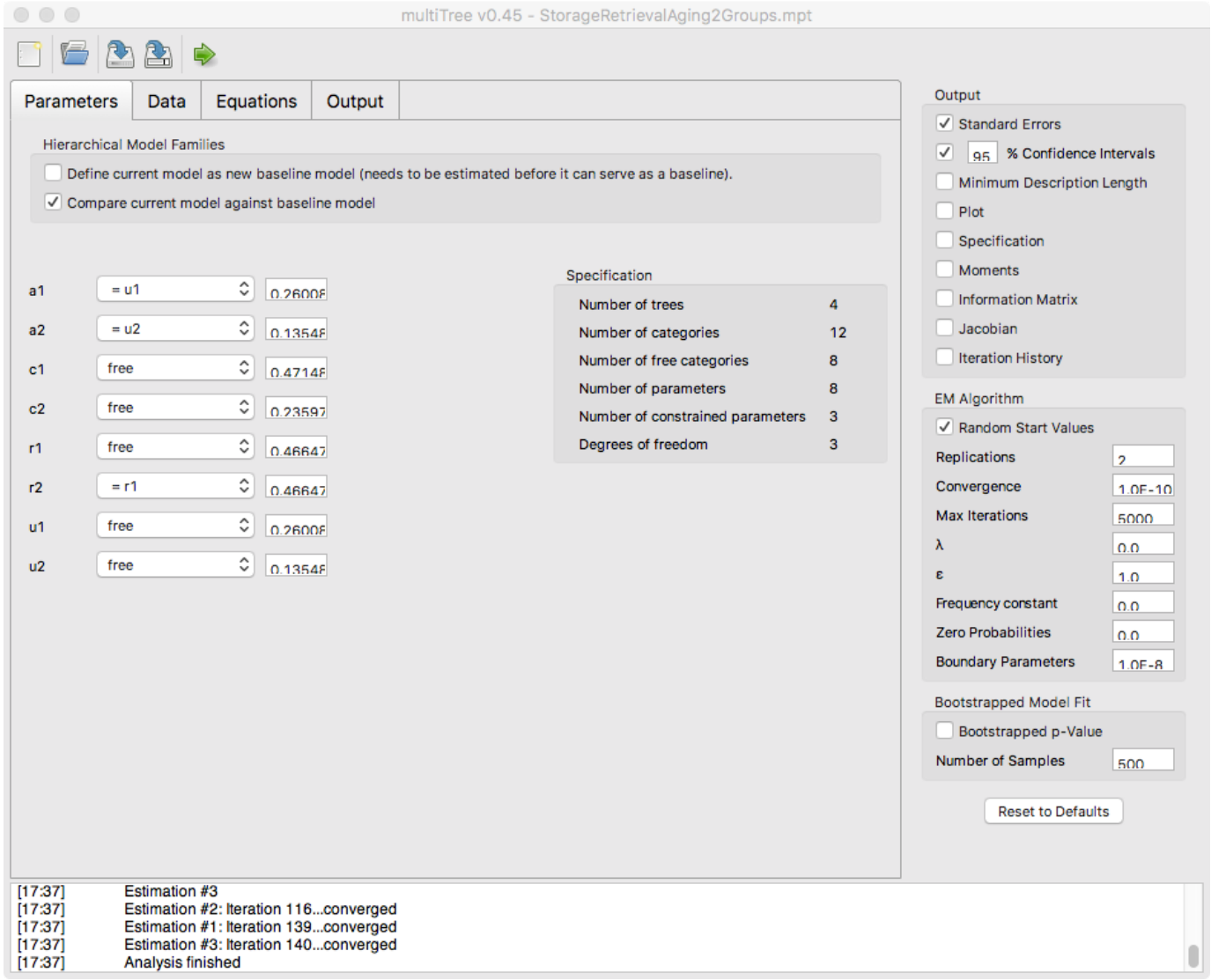
Example: Age group comparison



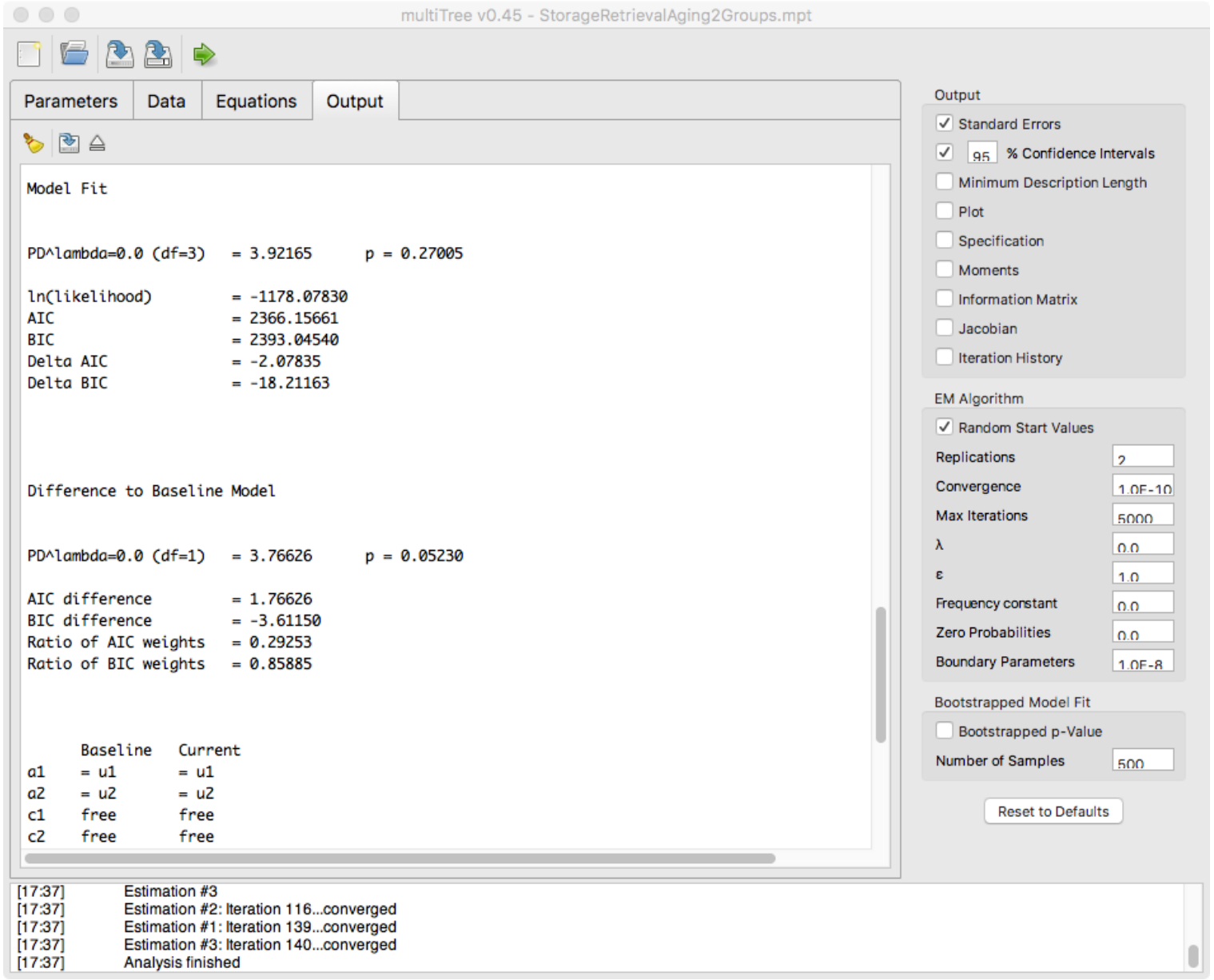
Definition of baseline model



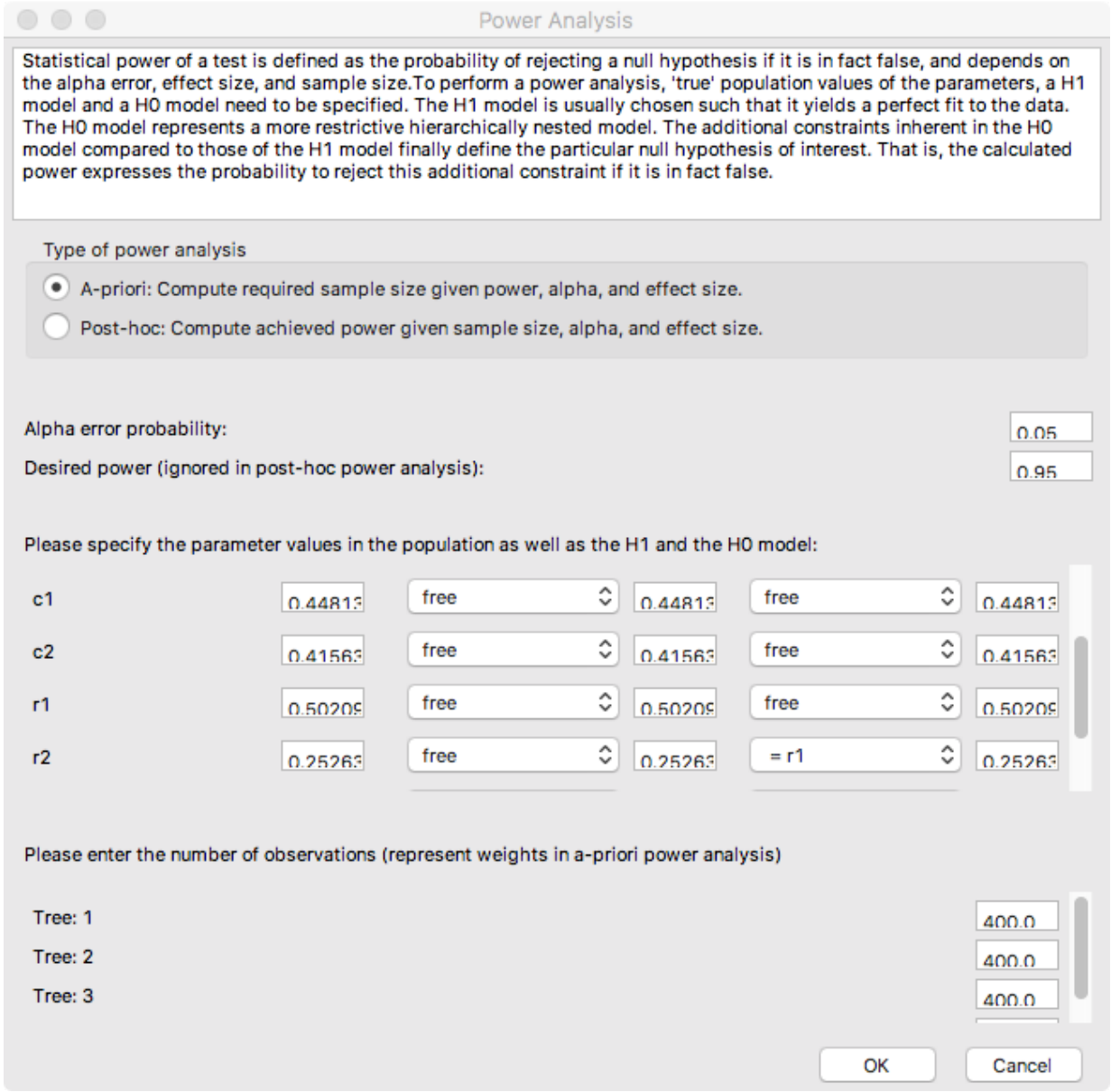
H_0: No age differences in retrieval



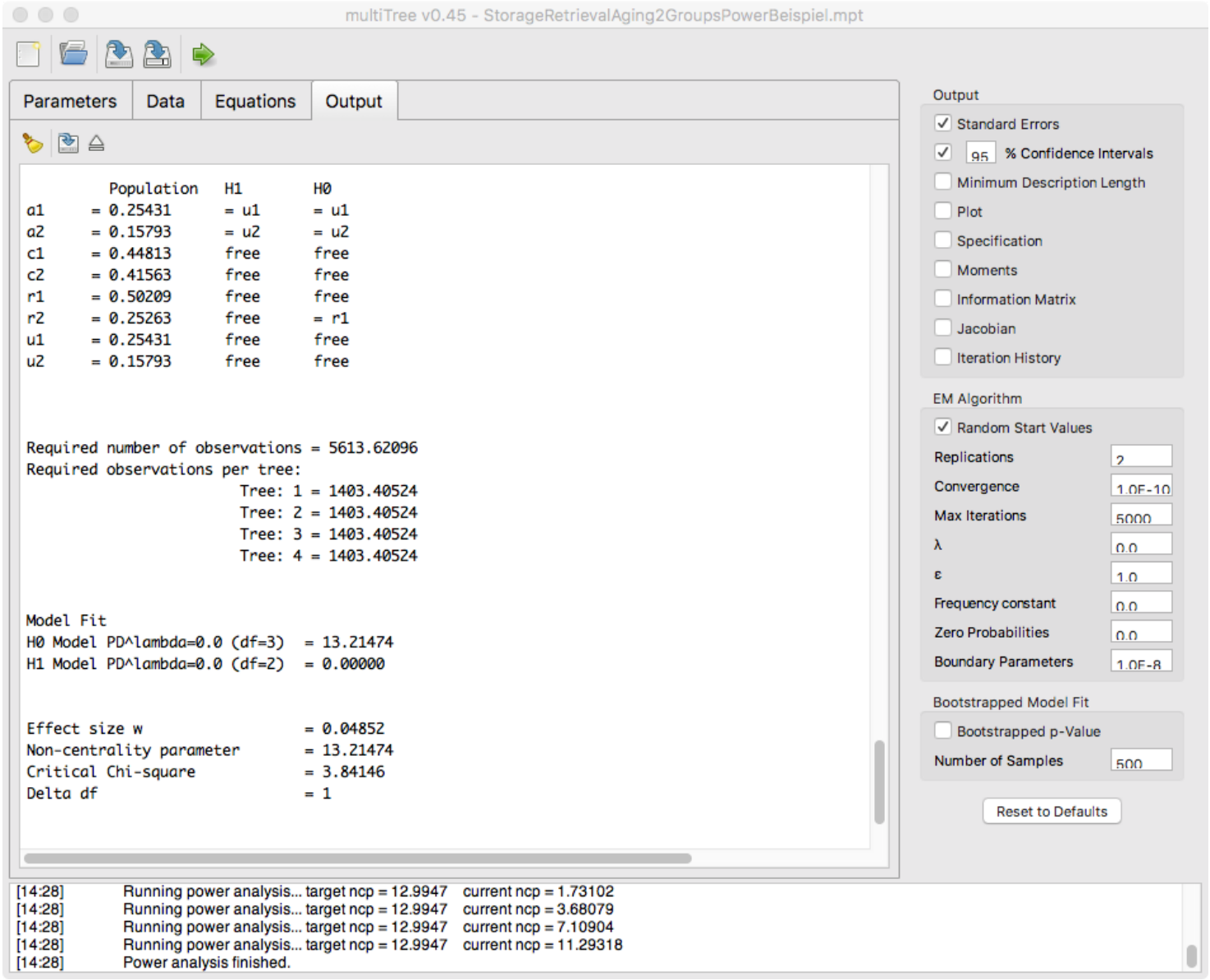
Is insignificant outcome due to lack of power?



multiTree power analysis window



multiTree power analysis output



3.3) Model selection

- Basic idea: Select the model that fits best *after correcting for model complexity/flexibility*
- Three Options:
- Penalty for *S*: Akaike Information Criterion (AIC)
  - $AIC(M_0) = -2 \cdot \ln(L(\boldsymbol{\theta};\mathbf{y})) + 2 \cdot S$
- Penalty for *S* taking *N* into account: Bayesian Information Criterion (BIC):
  - $BIC(M_0) = -2 \cdot \ln(L(\boldsymbol{\theta};\mathbf{y})) + S \cdot \ln(N)$
- Penalty for *S* taking *N* and complexity due to functional form into account: Normalized Maximum Likelihood (NML)

Normalized Maximum Likelihood (NML)

An implementation of the MDL principle was provided by Rissanen (2001) who derived the normalized maximum likelihood (NML) to measure the stochastic complexity of a model given a data set,

$$\text{NML} = -\text{LML} + C_{\text{NML}}(N), \tag{1}$$

where the maximum log-likelihood (LML) as a measure of fit is weighted against the complexity term

$$C_{\text{NML}}(N) = \ln \int_{\mathcal{X}^N} f(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}))d\mathbf{x}. \tag{2}$$

This complexity term is the natural logarithm of the integral over the maximum likelihoods across the whole outcome space $\mathcal{X}^N$ of potentially observable vectors $\mathbf{x}$ with N observations. Accordingly, a complex model that fits a wide range of observable data vectors will have a large value of $C_{\text{NML}}(N)$ compared to a model that fits only a small subset of observable data (Myung, Navarro, & Pitt, 2006). Unfortunately, there is no general closed-form expression of $C_{\text{NML}}(N)$ and numerical estimation techniques such as Monte Carlo (MC) integration are often too time intensive for practical purposes. An alternative is the Fisher information approximation (FIA; Rissanen, 1996),

$$\text{FIA} = -\text{LML} + C_{\text{FIA}}(N), \tag{3}$$

which is asymptotically equivalent to NML. The complexity term $C_{\text{FIA}}(N)$ covers the number of free parameters S and the number of observations N in the first summand and considers the functional form of the model in the second,

$$C_{\text{FIA}}(N) = \frac{S}{2} \ln \left(\frac{N}{2\pi} \right) + \ln \int_{\Omega} \sqrt{|\mathbf{I}(\boldsymbol{\theta})|} d\boldsymbol{\theta}, \tag{4}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix of sample size one. This

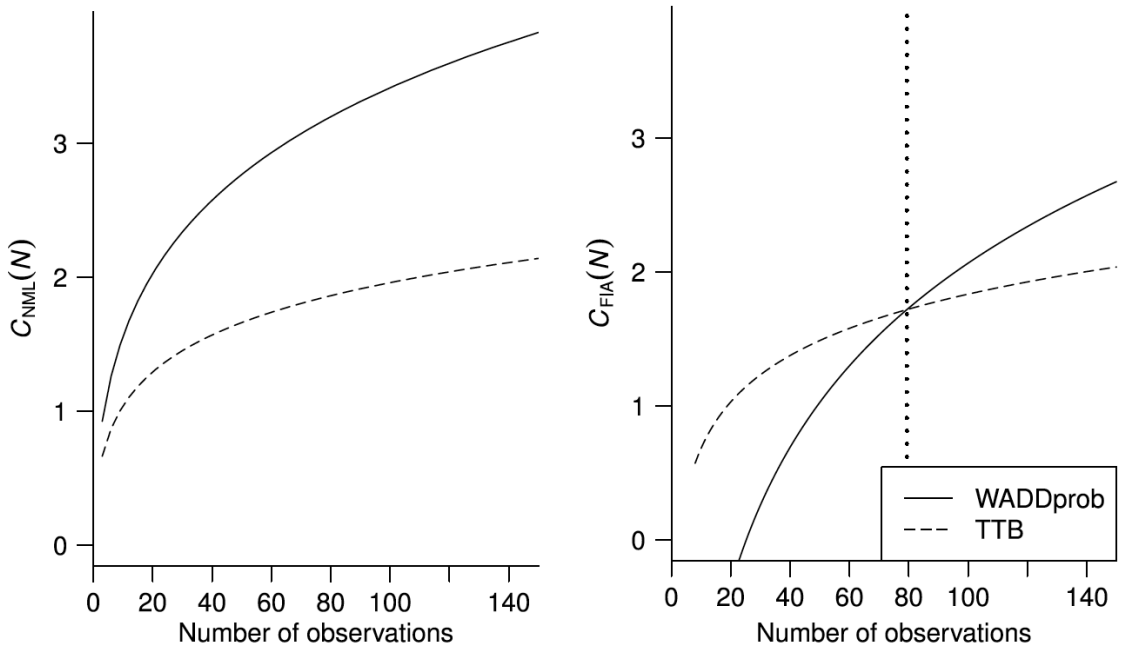
Heck, Moshagen & Erdfelder (2014)

To avoid biases in model selection using FIA for NML stable models, we propose to check whether the $C_{\mathrm{FIA}}(N)$ rank order of the candidate models is invariant across different numbers N of observations. Based on the definition of FIA in (4) it is easy to show that for any two models with a fixed but unequal number of parameters S_i and S_j, respectively, the $C_{\mathrm{FIA}}(N)$ rank order cannot be identical for all possible sample sizes. Since the integral in (4) is independent of N, it is straightforward to determine the (single) sample size $N'_{i,j}$ for which the complexity terms of two models with $S_i \neq S_j$ are equal. Equating the FIA terms of two models i and j and solving for N yields

$$N'_{i,j} = 2\pi \exp\left[\frac{2}{S_i - S_j}\left(\ln \int_{\Omega_{\mathbf{j}}} \sqrt{|\mathbf{I_j}(\boldsymbol{\theta})|}d\boldsymbol{\theta} - \ln \int_{\Omega_{\mathbf{i}}} \sqrt{|\mathbf{I_i}(\boldsymbol{\theta})|}d\boldsymbol{\theta}\right)\right]. \quad (5)$$

When $N > N'_{i,j}$, the $C_{\mathrm{FIA}}(N)$ terms of the two competing models i and j will always result in the same rank order. Because $C_{\mathrm{FIA}}(N)$ approximates $C_{\mathrm{NML}}(N)$ for increasing N, this must be the correct (i.e., NML-consistent) rank order. By implication, for any $N < N'_{i,j}$ the rank order of complexity terms is incorrectly inverted.

Heck, Moshagen & Erdfelder (2014)



(a) NML complexity. (b) FIA complexity.

Fig. 1. NML and FIA complexities for two decision strategies, take-the-best (TTB) and a probabilistic weighted-additive rule (WADDprob; Fig. 4), as functions of the number of observations N. The dotted vertical line for FIA marks the lower bound $N' = 80$.

3.4 Summary: Model Evaluation

- Check Model Identifiability
- Check (and optimize) statistical power of model test
- Goodness-of-fit test
- Model selection (several candidate models)
- Model Validation
- Model Application